Reference Libraries of Mass Spectra for Health and Environmental Science

The establishment of chemical identity is often the first and most difficult measurement task in a chemical analysis involving trace species. The most widely used method for making these identifications matches a measured mass spectrum against the reference mass spectrum of a known compound. While this method appears similar to traditional fingerprint analysis, the underlying data are expressions of fundamental molecular behavior. This enables the development and use of advanced quality control methods based on fundamental considerations.

S.E. Stein, Q.L. Pu, S. Yang, J. Roth (Div. 838)
C. Clifton (Div. 838 Guest Researcher),
P. Neta, L.E. Kilpatrick (Div. 838)

The ongoing NIST program in *Reference Libraries of Mass Spectra* is concerned with two varieties of mass spectra: those following ionization of gas-phase molecules by high-energy electrons (electron ionization, EI), and those originating from the collision-induced dissociation (CID) of gas-phase ions formed by "soft" methods, especially electrospray (ESI) and matrix assisted laser-desorption ionization (MALDI). ESI allows for large, non-volatile molecules to be analyzed directly from the liquid phase, while in MALDI gas-phase ions are formed from molecules that are present in the matrix.

Electron Ionization: The separation of volatile compounds in a gas chromatograph followed by the measurement of their spectra by EI mass spectrometry is widely used to reliably identify compounds at low concentrations. A key step is matching acquired spectra with reference spectra in a trusted library. NIST has long produced the most widely used comprehensive reference EI library, the NIST/EPA/NIH Mass Spectral Library, which is installed on thousands of instruments each year. Our ongoing work consists both of providing data for important new or newly identified compounds (for example, brominated pollutants that are byproducts of water treatment), expanding coverage of areas where analytical methodologies are changing (trimethyl silyl derivatives of metabolites), and continuous improvement by providing better data where evaluation has indicated that the data currently in the database can be improved.

NIST/EPA/NIH Mass Spectral Library with Search Program: (Data Version: NIST 05, Software Version 2.0d)

The most recent version of the NIST/EPA/NIH Mass Spectral Library of EI spectra was issued June 2005. It contains spectra and associated identification data for 163,198 compounds.

A gas-phase chromatography retention index collection of 25,727 compounds was introduced in this version to improve the confidence of identification by matching observed with reference retention values. For this version significant improvements in data consistency were made by using a NIST-developed chemical structure analysis method that identified thousands of compounds expected to have similar spectra (stereoisomers, tautomers). These were collectively evaluated to enhance reliability and quality of the data while minimizing redundancy.

Collision-Induced Dissociation Mass Spectra: With the relatively recent invention of mass spectrometers capable of measuring spectra for nonvolatile compounds, mass spectrometry has been extended to new application areas, most notably to the biosciences where most substances of interest are non-volatile. Data in this area have often had poor reproducibility. Our work is designed to develop standard data, which will allow far better reliability in compound identification from the CID spectra.

The use of CID mass spectra is becoming one of the most important areas in biomarker research. To support measurements this field, a new library of CID mass spectra for nearly 2,000 ions was added to the latest release of the Mass Spectral Library.

These spectra were measured at NIST, by collaborators, and extracted from the literature, and have undergone careful evaluation. Addition of these spectra required major changes to the structure of data archive systems due to a significant increase in "metadata" for these spectra. Quality control methods were modified to deal with the inherent variability of energy, hence spectral patterns, of this class of mass spectra.

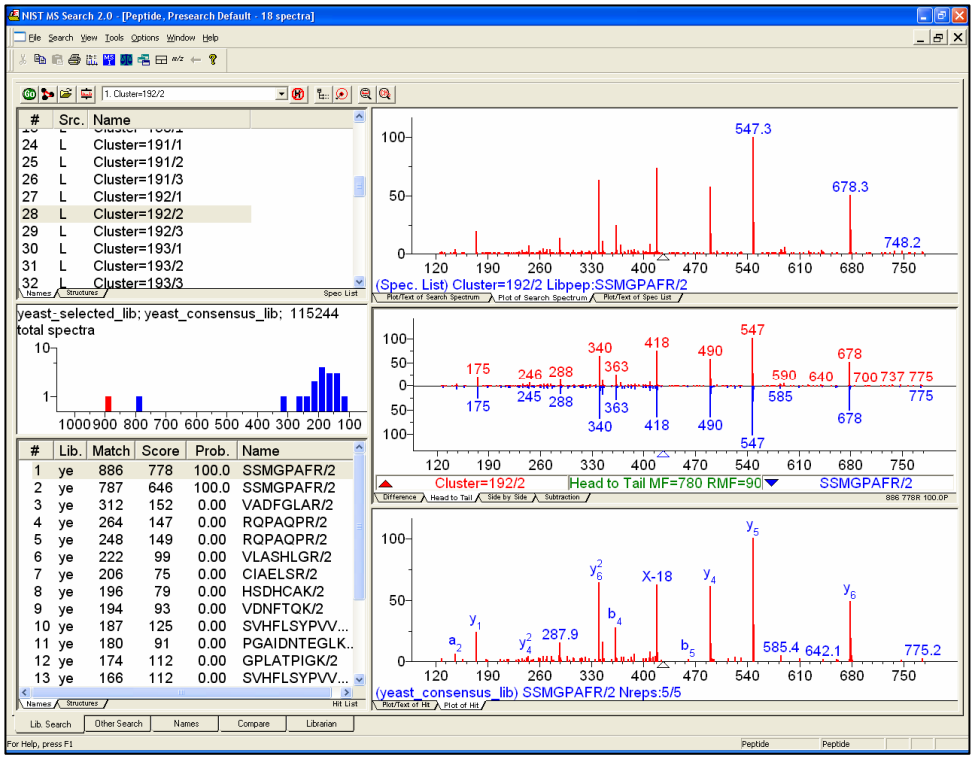
We have recently begun to build a library of peptide CID spectra. In the area of "proteomics" peptide spectra often serve as the fundamental unit of measurement. They serve to "sequence" peptides and thereby identify their parent proteins. In building the library, most spectra originate from data files provided by other laboratories where complex mixtures of proteins were digested (converted to "tryptic" peptides) and analyzed by LC-MS/MS using ESI. Peptide identifications are made at NIST using sequence-matching search engines, with reliabilities determined by a series of quality control methods. This has involved collaboration with a number of laboratories concerned with biological aspects of these measurements.

Quality control for the large number of spectra observed in these experiments requires automated processing methods. These have been developed, refined, and are now being used to separate true and false positive spectra and to better measure reliabilities. Other studies deal with the variability of mass spectra that originate from energetic and structural effects. Further studies are extending search algorithms for EI spectra to CID peptide spectra. This has also led to the development of an experimental program to measure spectra of well-defined peptides under carefully controlled condition for both library building and to aid quality control methods.

In collaboration with the Institute of Systems Biology in Seattle, a first library has been developed for yeast, a particularly well-studied organism in proteomics. This contains a searchable library that is fast, and contains over 35,000 different peptide ions derived from thousands of data files acquired in many laboratories.

INSTITUTE FOR Systems Biology *Revolutionizing science. Enhancing life.*

Tests show that the peptide identification method developed generates several-fold more identifications than current sequence-search methods at speeds one hundred times faster. The library is currently under testing in several laboratories and is being refined with the plan to release the public version in 2006. Also, under development is a library of peptides derived from human proteins which is expected to find broad application in the field of biomarker discovery.



Impact: Speed and accuracy of analysis is a vital issue in many – perhaps most – cases. Monitoring for dangerous toxins, chemical weapons, food contamination, and health/disease status biomarkers is an enormous challenge with extremely high reward in terms of quality of life and security for the nation.

Future Plans: Libraries of peptide ion spectra targeted at a specific organism or biological system are in the early stage of development, and testing and their efficacy will need to be ascertained. A number of research studies to understand systematic similarities and difference in peptide fragmentation, to further enhance analysis algorithms, are also planned.